面向机器学习模型的基于 PCA 的成员推理攻击

彭长根[1,2,3]，高婷[1,2]，刘惠篮[1]，丁红发[3,4]

（1. 贵州大学公共大数据国家重点实验室，贵州 贵阳 550025；2. 贵州大学密码学与数据安全研究所，贵州 贵阳 550025；
3. 贵州大学计算机科学与技术学院，贵州 贵阳 550025；4. 贵州财经大学信息学院，贵州 贵阳 550025）

摘 要：针对目前黑盒成员推理攻击存在的访问受限失效问题，提出基于主成分分析（PCA）的成员推理攻击。首先，针对黑盒成员推理攻击存在的访问受限问题，提出一种快速决策成员推理攻击 fast-attack。在基于距离符号梯度获取扰动样本的基础上将扰动难度映射到距离范畴来进行成员推理。其次，针对快速决策成员推理攻击存在的低迁移率问题，提出一种基于 PCA 的成员推理攻击 PCA-based attack。将快速决策成员推理攻击中的基于扰动算法与 PCA 技术相结合来实现成员推理，以抑制因过度依赖模型而导致的低迁移行为。实验表明，fast-attack 在确保攻击精度的同时降低了访问成本，PCA-based attack 在无监督的设置下优于基线攻击，且模型迁移率相比 fast-attack 提升 10%。

关键词：机器学习；对抗样本；成员推理攻击；主成分分析；隐私泄露

中图分类号：TP309.7

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022009

PCA-based membership inference attack for machine learning models

PENG Changgen[1,2,3], GAO Ting[1,2], LIU Huilan[1], DING Hongfa[3,4]

1. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
2. Institute of Cryptography and Data Security, Guizhou University, Guiyang 550025, China
3. College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
4. College of Information, Guizhou University of Finance and Economics, Guiyang 550025, China

Abstract: Aiming at the problem of restricted access failure in current black box membership inference attacks, a PCA-based membership inference attack was proposed. Firstly, in order to solve the restricted access problem of black box membership inference attacks, a fast decision membership inference attack named fast-attack was proposed. Based on the perturbation samples obtained by the distance symbol gradient, the perturbation difficulty was mapped to the distance category for membership inference. Secondly, in view of the low mobility problem of fast-attack, a PCA-based membership inference attack was proposed. Combining the algorithmic ideas based on the perturbation category in the fast-attack and the PCA technology to suppress the low-migration behavior caused by excessive reliance on the model. Finally, experiments show that fast-attack reduces the access cost while ensuring the accuracy of the attack. PCA-based attack is superior to the baseline attack under the unsupervised setting, and the migration rate of model is increased by 10% compared to fast-attack.

Keywords: machine learning, adversarial example, membership inference attack, principal component analysis, privacy leakage

收稿日期：2021−10−14；**修回日期**：2022−01−05

基金项目：国家自然科学基金资助项目（No.U1836205, No.62002080）；贵州省科技计划基金资助项目（黔科合平台人才[2020]5017）；贵州省教育厅自然科学基金资助项目（黔教合 KY 字[2021]140）；贵州大学人才引进科研基金资助项目（贵大人基合字[2020]61）

Foundation Items: The National Natural Science Foundation of China (No.U1836205, No.62002080), The Science and Technology Plan Foundation of Guizhou Province (No.[2020]5017), The Natural Science Foundation of Department of Education of Guizhou Province (No.[2021]140), The Research Project of Guizhou University for Talent Introduction (No.[2020]61)

0 引言

物联网、大数据、云计算等新兴技术使海量数据的采集、存储和处理成为可能，人工智能特别是机器学习理论与技术的快速发展，使其在安防、交通、医疗等各领域得到了广泛应用。与此同时，机器学习的安全与隐私问题成为人们关注的焦点，有学者提出了对抗样本攻击[1]、数据投毒攻击[2]、模型推断以及成员推理[3-4]等各类安全与隐私攻击模型。这些有效的攻击方法引发了人们对机器学习的担忧，同时也成为机器学习发展的内生动力之一，推动科学研究人员和工程技术人员研发安全性与隐私性更好的机器学习算法和模型[5]。研究机器学习隐私攻击模型能够推动人们更加深入地理解机器学习模型的深层机理，揭示隐私泄露的本质原因，有利于更好地防范机器学习模型的隐私泄露风险，并有利于推动设计更加高效保护隐私的机器学习模型。

机器学习成员推理攻击是敌手通过分析机器学习模型来推断目标数据样本是否包含于该机器学习模型训练样本数据集的一种隐私攻击方法，该攻击主要作用于训练样本数据集，威胁机器学习训练样本的成员关系隐私。现有工作大致可分为黑盒成员推理攻击和白盒成员推理攻击两类。

在黑盒成员推理攻击中，一类方法是基于模型预测置信度的成员推理[3-4,6-7]；另一类方法是基于标签决策的成员推理[8-10]。这两类攻击方法仅能通过查询目标模型获得输入输出对，而不能获得任何关于模型的额外数据，即借助目标模型的输出结果来完成成员推理。其中，基于模型预测置信度的成员推理作为一种需要借助目标模型的置信向量输出来进行推断的技术，能够实施成功源于机器学习固有的过拟合特性，即成员数据的输出向量的分布更集中，而非成员数据的输出向量的分布相对平缓。尽管这些工作在黑盒设置下取得了不错的进展，但由于企业的访问限制，敌手无法从目标模型中获得足够多样本的预测向量。更关键的是，这类攻击模型难以突破 MemGuard[11]防御。因此，研究者进一步提出基于标签决策的成员推理，其仅需借助目标模型的输出标签即可进行成员推理，推断者将模型返回的最大预测标签作为推断输入，在预测模型训练集与测试集的过程中引入了扰动难度，提高了成员推理的稳健性，因此被广泛应用于机器学习的安全和隐私领域。预测标签与对抗样本、影子技术[3]相结合，能够提升模型的稳健性及推理精度，但其难以保证推理的可信度和数据访问的低成本与可迁移性。例如，Yeom 等[8]定量分析了训练集和测试集的攻击性能与损失之间的关系，提出了基于过拟合特性下的基线攻击。随后，Choo 等[9]提出了一种类似边界攻击的方法。通过将机器学习的过拟合特性映射到训练集样本与测试集样本的扰动问题中，借助对抗样本解决传统成员推理固有的过拟合问题。但是，该类攻击访问成本过，限定访问次数会导致攻击失效，这在一定程度上削弱了算法的推断精度，给推断者的具体实施带来了巨大挑战。

在白盒成员推理攻击中[12-15]，攻击者可以对目标模型进行白盒访问。在此条件下，攻击者可以获得目标模型所使用的云训练平台的相关信息，或直接获得目标模型的训练算法、内部参数、模型结构、中间结果等信息，从而构建与目标模型预测能力相似的模型。鉴于先前的攻击方法很少用到这些信息，于是，Nasr 等[12]将成员推理攻击拓展到基于先验知识的白盒设置，将从目标模型获得的激活函数和梯度信息作为推断的特征，来进行成员推理，还提出了针对联邦学习的主动成员推理攻击。接着，Hayes 等[13]在应对生成对抗网络（GAN, generative adversarial network）的成员推理攻击的工作中也提到了一种白盒攻击，该攻击仅使用 GAN 鉴别器部分的输出，而不需要鉴别器或生成器的学习权重即可完成推断。除此之外，Long 等[15]提出了一种针对泛化性良好的模型的成员推理攻击并称为 GMIA。在此种模型下，不是所有的数据都易遭受成员推理攻击，因此需要找到易受到成员推理攻击的脆弱数据点来进行推理。尽管现有的白盒成员推理能够实现较好的攻击效果，但由于在实际场景中机器模型通常部署为黑盒模型，其所需的模型知识在实际机器学习应用场景中难以得到满足。

综上，黑盒成员推理攻击在机器学习模型中有更加广泛的应用，但现有的黑盒成员推理攻击存在访问成本高、可迁移性弱、稳健性差等问题。针对这些问题，本文通过引入决策边界搜索过程中基于距离的符号梯度方法[16]，从扰动样本出发将扰动难度映射到距离范畴，提出一种快速决策成员推理攻击 fast-attack。其次，针对快速决策成员推理攻击存在的低迁移率问题，将快速决策成员推理攻击中的基于扰动算法与主成分分析（PCA, principle component analysis）技术相结合，本文提出一种基于

PCA 的成员推理攻击 PCA-based attack，以抑制 fast-attack 因过度依赖模型而导致的低迁移行为。本文的具体贡献如下。

1) 提出一种快速决策成员推理攻击 fast-attack。以预测标签作为模型的输入，通过引入自适应贪婪算法与二分搜索来确定决策边界的对抗样本，将扰动难度映射到距离范畴来寻找预测差异，从而实现成员推理，降低了攻击参与方的查询成本，适用于低成本攻击的目标场景。

2) 提出一种基于主成分分析的成员推理攻击 PCA-based attack。基于流形界面对高维数据的影响设计基于主成分技术的嵌入映射，通过逻辑判别实现细粒度的成员推理，解决了 fast-attack 因过度依赖模型造成的过拟合特定机器学习模型的问题。

3) 仿真实验表明，fast-attack 在降低访问成本的同时攻击精度达到 75%。而 PCA-based attack 在无监督的设置下优于基线攻击，攻击性能与目前黑盒成员推理攻击相匹敌，且模型迁移率比 fast-attack 提升 10%。除此之外，还评估了 2 种算法的抵抗防御能力，实验表明本文攻击对大多数防御技术都具有不错的攻击效果，且具有强稳健性。

1　基础知识

本节主要介绍成员推理攻击涉及的数学符号和相关定义。

1.1　符号说明

本文所涉及数学符号如表 1 所示。

表 1　　**符号说明**

符号	定义	符号	定义
f	目标模型	Ω	先验知识
h	推断模型	$\mathcal{A}$	成员推理
H	流形模型	c	样本标签
x	目标数据	L	损失函数
x^*	映射数据	n	扰动像素点数目
x_{adv}	对抗样本	g	映射方向
η	扰动步长	S'_{adv}	相交区域
τ	判别阈值	ϕ	单调函数
$d(x,x_{\text{adv}})$	扰动难度	$\boldsymbol{P}$	特征向量
S_{AR}	对抗区域	δ	扰动大小
Q	访问量	$\delta_{\max}$	最大扰动量
$Q_{\max}$	最大访问量	λ	迭代步长
f_1	分类器 1	f_2	分类器 2

1.2　相关定义介绍

1.2.1　成员推理攻击

成员推理攻击是一种通过分析目标模型来确定给定数据样本是否存在于该目标模型的训练集中的攻击方法[3]。当给定 x，目标模型 f 以及敌手的先验知识 Ω，得到相应的成员推理攻击为

$$\mathcal{A}:\ x,f,\Omega \to \{-1,1\} \tag{1}$$

其中，1 代表 x 存在于目标模型的训练数据集中，反之不存在。

1.2.2　流形学习

流形学习是一种新的机器学习方法，它能够对训练集中的高维数据空间进行非线性降维，揭示其流形分布，从中找到隐藏在高维观测数据中有意义的低维结构，以便从中提取易于识别的特征。其目标是发现嵌入高维数据空间中的低维流形结构，并给出一个有效的低维表示。

1.2.3　主成分分析

主成分分析是一种线性数据变换方式，可以把可能具有相关性的高维变量合成线性无关的低维变量，数据在主成分方向上的投影拥有最大方差。该技术的主要目标是通过线性变换寻找一组最优的单位正交向量基，并用它们的线性组合来重构原样本，以使重构后的样本和原样本的误差最小。

2　成员推理攻击

基于训练样本比测试样本更难被扰动的假设原理，本文提出新的成员推理攻击，其流程如图 1 所示。

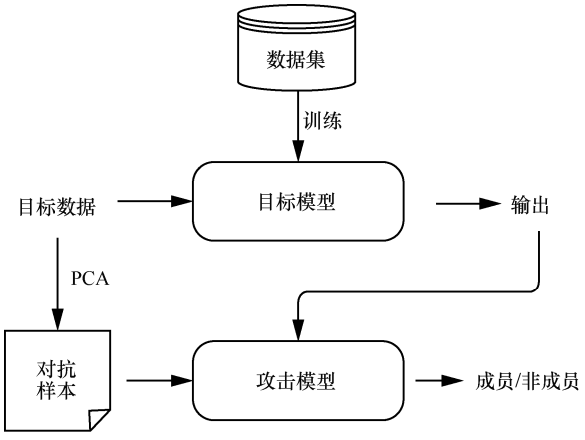


图 1　成员推理攻击的流程

由图 1 可知，给定目标数据，通过分析目标模型得到相应的预测标签。结合目标数据和预测标签

作为攻击模型的输入，得到成员与非成员的决策判别。其中，攻击模型的设定通过确定决策边界，将扰动难度映射到距离范畴来构建快速决策成员推理攻击。此外，将前者基于扰动的攻击方案与主成分分析技术相结合，不需要访问目标模型，进一步构造出基于主成分分析的成员推理攻击。与一般的成员推理攻击方法不同，本文借助扰动难度来区分成员样本和非成员样本，该扰动难度主要通过目标数据与其对抗样本之间的欧氏距离来衡量，实现以较少先验信息资源高效推断出目标模型的训练数据集隐私数据，降低成本需求。

2.1 快速决策成员推理攻击

针对目前大多数黑盒成员推理攻击因过拟合而导致的高精度攻击这一问题，以及目前基于标签决策的成员推理存在的高反馈访问成本问题，本文引入文献[16]的扰动样本生成方案，构造了一个快速决策成员推理攻击 fast-attack。该攻击主要包含 2 个步骤：对抗样本生成和逻辑判别。首先以预测标签作为模型的输入，引入自适应贪婪算法与二分搜索对目标进行决策变动，生成对抗样本；然后计算对抗样本与原始目标之间的欧氏距离，将扰动难度映射到距离范畴来寻找目标模型的训练数据和测试数据的预测差异；最后将预测差异进行逻辑判别获得细粒度的成员信号，以实现目标人群的成员推理。

通过将机器学习的过拟合特性映射到训练集样本与测试集样本的扰动问题中，借助对抗样本解决传统成员推理固有的过拟合问题。通过将自适应贪婪算法与二分搜索相结合来确定决策边界，解决了目前黑盒成员推理攻击固有的高成本问题。

在对抗样本生成的过程中，首先通过向源数据添加高斯扰动得到对抗样本的初始值，然后引入二分搜索和自适应贪婪算法沿着对抗性区域和非对抗性区域之间的边界执行随机游走，使它停留在对抗区域，并且减小到目标图像的距离。最后，结合获得的扰动样本来提取关于分类器决策边界的细粒度信息，从而进行成员推理。

定义 1 对抗样本生成中得到的损失函数为

$$\begin{aligned} &L(x,\delta)=\sum(x_{\text{adv}}-x)^2\\ &\text{s.t. } x_{\text{adv}}=x+\delta\\ &c(x)\neq c(x_{\text{adv}})\\ &\|\delta\|\leqslant\delta_{\max},Q\leqslant Q_{\max} \end{aligned} \tag{2}$$

其中，$c(x):=\arg\max_{i\in[k]} f(x)_i$ 为机器学习的样本标签。进一步化简得

$$\begin{aligned} &\text{minimize } L(x,\delta)=\|\delta\|_p+au(x+\delta)\\ &\text{s.t.}\quad \|\delta\|\leqslant\delta_{\max},Q\leqslant Q_{\max}\\ &x+\delta=\min(\max(x+\delta,0),1) \end{aligned} \tag{3}$$

其中，$u(\cdot)=\min(\max_{i\neq t} f_i(\cdot)-f_t(\cdot),0)$。

该损失函数计算是一个难解问题，因此，本文基于贪婪算法的局部随机优化进行边界搜索，得到映射方向 g 为

$$\begin{aligned} &\text{sgn}(\nabla_\delta L(x,\delta))=g(\overrightarrow{x,x_t})\\ &\|\text{sgn}(\nabla_\delta L(x,\delta))\|=d(x,x_t)\\ &d(x,x_t)-d(x,x_{t+1})=\epsilon d(x,x_t)\\ &\epsilon>0,g=1 \end{aligned} \tag{4}$$

其中，ϵ 为方向距离参数，x_t,x_{t+1} 为迭代扰动点。接着，沿着该方向以一定步长进行随机边界游走，多次迭代搜索生成相应的对抗样本为

$$x_{t+1}=x_t+\lambda\,\text{sgn}(\nabla_\delta L(x,\delta)) \tag{5}$$

最后，计算对抗样本与原始目标数据之间的欧氏距离 $L_p(x,x_{\text{adv}})$，并与获得的阈值 τ 进行判别完成攻击。具体如下，给定目标数据点到模型边界的距离的估计 $\text{dist}_f(x)=\min\|x-x_{\text{adv}}\|_p$。如果 $\text{dist}_f>\tau$，则将 x 分类为训练集成员。如果 $\text{dist}_f=0$，则认为该目标数据点在决策边界上，分类错误。同时调整阈值 τ，使该算法在本实验数据上效果最佳。

综上，fast-attack 的伪代码如算法 1 所示。

算法 1 fast-attack

输入 目标数据集 $\mathcal{D}$，参考样本点 x_{re}，扰动像素点数目 n，噪声 θ，最大扰动量 $\delta_{\max}$，迭代步长 λ，距离范数 p，阈值 τ，最大访问量 $Q_{\max}$

输出 对抗样本 x_{adv}，成员推理 m

1) $x\in\mathcal{D}$，结合 x_{re} 利用二分法在对抗区间搜索，更新 $x_{\text{adv}}=x_i$

2) while $(d(x,x_{\text{adv}})\leqslant\delta_{\max})\,\&\,(Q\leqslant Q_{\max})$

① 随机选取 x_{adv} 中的 n 像素点进行高斯随机扰动，得到 x'，计算 $x_g=x_i+\theta x'$

结合 x_g，利用二分法得到新的 x_g

if $d(x,x_g)>d(x,x_i)$

direction= −1

end

direction= 1

② 初始化 $x_e = x_g$

while $(d(x,x_e) > d(x,x_i)) \,\&\, (\lambda > 0.001)$

$\Delta = \text{direction}(x_g - x_i)$

$x_e = x_i + \lambda\Delta$

$\lambda = \lambda / 2$

③ 借助 x_{re}，利用二分法得到 x_{adv}

3) 得到对抗样本 x_{adv}

4) 最后，计算 $\text{dist}_f(x) = \min\left\|x - x_{\text{adv}}\right\|_p$

if $\text{dist}_f(x) > \tau$

$m = 1$

end

$m = -1$

5) 返回 x_{adv}, m

其中，步骤 1)是相关变量初始化；步骤 2)中的①保证在给定最大扰动及最大访问的条件下，借助自适应贪婪算法获得局部最优方向，使每个样本点接近决策边界；步骤 2)中的②、③表示沿着最优方向，进行迭代更新，获取最贴近决策边界的对抗样本点；步骤 3)、步骤 4)借助对抗样本进行逻辑判别，进而成功推断出目标样本点。算法中相关参数的取值见实验部分。

2.2　基于 PCA 的成员推理攻击

尽管上文提到的快速决策成员推理攻击能够降低模型交互产生的成本，但是面对访问受限、标记训练样本不足的系统，该种攻击将失去威胁效用。除此之外，该种攻击因过度依赖模型将导致攻击的迁移率低下。因此，本节针对以上问题提出一种新的改进攻击，即基于主成分分析的成员推理攻击 PCA-based attack，其将快速决策成员推理攻击中基于扰动算法与主成分分析技术相结合来完成成员推理，框架如图 2 所示。该攻击通过主成分分析技术模拟流模型生成对抗性区域，借助对抗性区域来构建决策区间进而实行成员推理，实现以较少先验信息资源有效推断出目标系统隐私数据，从而降低对目标系统历史访问信息的要求。

基于 PCA 技术，本文的成员推理攻击可划分为以下 3 个阶段。

1) 对抗区域生成阶段

尽管已有的成员推理攻击对泛化性能良好的模型[17-20]失效，但广义良好的模型对分布在 x 点与流形切平面正交方向上的畸变高度敏感。成员推理中，需要寻找成员与非成员数据的识别特征差异，进而实行判别。数据的识别特征差异可以通过非线性降维，揭示其流形分布，从中找到隐藏在高维观测数据中有意义的低维结构，以便从中提取易于识别的特征。因此，在这一阶段通过 PCA 技术进行数据降维，在低维流形界面[21]寻找数据的正交映射方向，并选取满足条件的扰动步长，最终获取原始数据的对抗区域。

定义 2　流形界面为 H，流形界面的映射样本点为

$$z_H(x) := \underset{z}{\text{argmin}} \left\|x - H(z)\right\|_p \tag{6}$$

其中，映射点 $x_0^* = H(z_H(x_0))$。

沿用 Zhang 等[22]的定义，得到对抗区域为

$$S_{\text{AR}}(x_0) := \left\{ x \,|\, x = x_0 + \eta \frac{x_0 - x_0^*}{\left\|x_0 - x_0^*\right\|_2}, \eta \in [\eta_l, \eta_u] \right\} \tag{7}$$

其中，η_l 表示最小的误分类扰动步长，η_u 表示不易察觉的最大扰动步长。

2) 对抗样本生成阶段

由于对抗区域依赖于独立于分类模型的数据流形，因此可以根据对抗区域的定义，用无监督方法生成对抗性示例。计算过程为

$$x_{\text{adv}} := x + \eta \frac{x - x^*}{\left\|x - x^*\right\|_2} \tag{8}$$

其中，流形 M 是很难显式构造的，特别是对于复杂的现实世界数据集。因此，投影点 $x^* = H(z_H(x))$ 不能直接计算。本文使用 PCA 技术来近似流形 M，以产生对抗性示例。推导过程为

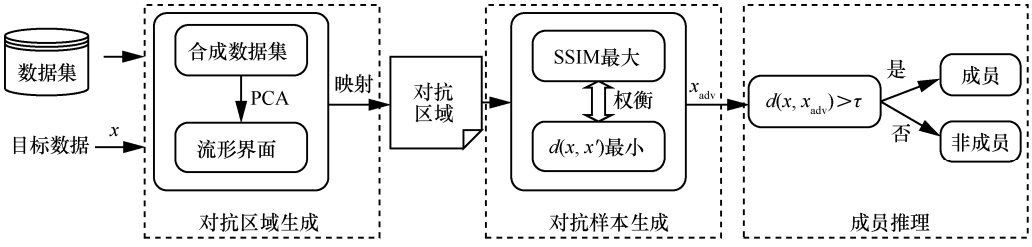


图 2　PCA-based attack 框架

$$x_{\text{adv}} := x+\eta\frac{x-x^*}{\left\|x-x^*\right\|_2}=$$
$$\sum_{i=1}^{k}p_i\left[\left(1+\frac{\eta}{\left\|x-x^*\right\|_2}\right)y_{i-}\frac{\eta}{\left\|x-x^*\right\|_2}y_i^*\right]+$$
$$\sum_{i=k+1}^{D}p_i\left[\left(1+\frac{\eta}{\left\|x-x^*\right\|_2}\right)y_{i-}\frac{\eta}{\left\|x-x^*\right\|_2}y_i^*\right]=$$
$$\sum_{i=1}^{k}p_iy_i+\sum_{i=k+1}^{D}p_i\left[y_i+\frac{\eta y_i}{\left\|x-x^*\right\|_2}-\frac{\eta y_i^*}{\left\|x-x^*\right\|_2}\right]=$$
$$x+\sum_{i=k+1}^{D}\frac{\eta}{\left\|x-x^*\right\|_2}p_iy_i\left(1-\frac{y_i^*}{y_i}\right) \tag{9}$$

其中，$\boldsymbol{Y}=\boldsymbol{P}^{\mathrm{T}}x$ 和 $\boldsymbol{Y}^*=\boldsymbol{P}^Tx^*$ 用于主成分降维来模拟流形界面，得到 $y_i^*=g(y_i)$ 且

$$g(y_i)\approx g(0)+g'(0)y_i+\frac{g''(0)}{2!}y_i^2 \tag{10}$$

故

$$\left|\phi(i)\right|=\left|1-\frac{y_i^*}{y_i}\right|=\left|1-\frac{g(0)}{y_i}-\frac{g''(0)}{2!}y_i-g'(0)\right| \tag{11}$$

为递减函数，可用简单函数替换。其中，使用结构相似性和距离最小化原则选取合适的扰动步长为

$$\eta := \alpha\underset{\eta_1}{\operatorname{argmax}}\,\text{ssim}(x,x_{\text{adv}})+(1-\alpha)\underset{\eta_2}{\operatorname{argmin}}d(x,x_{\text{adv}}) \tag{12}$$

3) 成员推理阶段

定义 3 成员推理函数 $h(x)$。用 $h(x)$ 表示目标数据是否存在于推断系统的训练集中，在逻辑判别函数的基础上，采用以下成员推理函数

$$S(x)=\min d(x,x_{\text{adv}})-\tau,\quad d(x,x_{\text{adv}})=\left\|x-x_{\text{adv}}\right\|_p \tag{13}$$

$$h(x)=\text{sign}(S(x))=\begin{cases}1, S(x)\geqslant 0\\ -1, \text{其他}\end{cases} \tag{14}$$

其中，$h(x)$ 为 1 时，代表 x 在目标模型的训练集中，反之不在。

综上所述，PCA-based attack 的伪代码如算法 2 所示。其中，步骤 1)对数据进行主成分降维处理，获得流形界面；步骤 2)～步骤 4)进行投影方向的搜寻，以获得流形界面的投影点，进而生成对抗样本；步骤 5)将原始数据与对抗数据之间的扰动距离通过阈值判别，进行成员推理。

算法 2 PCA-based attack

输入 合成数据集 $\mathcal{D}$，主成分参数 k，扰动步长 η，递减函数 $\phi(i)$，距离范数 p，判别阈值 τ

输出 对抗样本 $\mathcal{D}_{\text{adv}}$，成员推理 m

1) $x\in\mathcal{D}$，利用 PCA 计算 x 的降维数据 Y, $Y=$ PCA_transform(x)

2) 借助递减函数 $\phi(i)$ 划分降维数据点 Y

if $i>k$

$\hat{y}_i=\phi(i)y_i$

end

$\hat{y}_i=y_i$

3) 生成扰动 $\varepsilon=\eta\frac{x-x^*}{\left\|x-x^*\right\|_2}$，其中 $x^*=$ PCA_inverse(Y^*)

4) 生成对抗样本 $x_{\text{adv}}=x+\varepsilon$

5) 得到 $\mathcal{D}_{\text{adv}}$，计算 $\text{dist}_f(x)=\min\left\|x-x_{\text{adv}}\right\|_p$

if $\text{dist}_f(x)>\tau$

$m=1$

end

$m=-1$

6) 返回 $\mathcal{D}_{\text{adv}}, m$

3 方案分析

3.1 可行性分析

机器学习模型在预测训练集样本时能以更高的精准度进行预测。在过拟合的情况下，训练集样本的预测置信度明显高于测试集样本。因此可以判定训练集样本相比测试集样本更难被扰动。另外，针对二进制逻辑回归模型的特殊情况，给定学习权重向量 $\boldsymbol{\omega}$ 和偏置 b，逻辑回归模型的输出为判别类的置信向量

$$\boldsymbol{z}(x):=\sigma(\boldsymbol{\omega}^{\mathrm{T}}x+b) \tag{15}$$

其中，$\sigma(t)=\frac{1}{1+\mathrm{e}^{-t}}\in(0,1)$ 为逻辑函数。

该模型表明，点 x 的置信度与从 x 到模型决策边界的欧氏距离之间存在一定的正向关系。即从 x 到模型边界的距离为

$$\frac{(\boldsymbol{\omega}^{\mathrm{T}}x+b)}{\left\|\boldsymbol{\omega}\right\|_2}=\frac{\sigma^{-1}(\boldsymbol{z}(x))}{\left\|\boldsymbol{\omega}\right\|_2} \tag{16}$$

因此，获得点到边界的距离所产生的信息与已知模型的预测置信度的效果相同。部分研究表明[19-21]，

成员推理的实施可通过计算目标点到边界的距离，而其正是找到最小对抗性扰动的问题。

对验证数据进行实验分析（见附录 1），决策判别如图 3 所示，成员样本相比于非成员样本，距离决策边界更远，更难被扰动，进一步说明成员推理攻击可转变为求最小扰动问题。

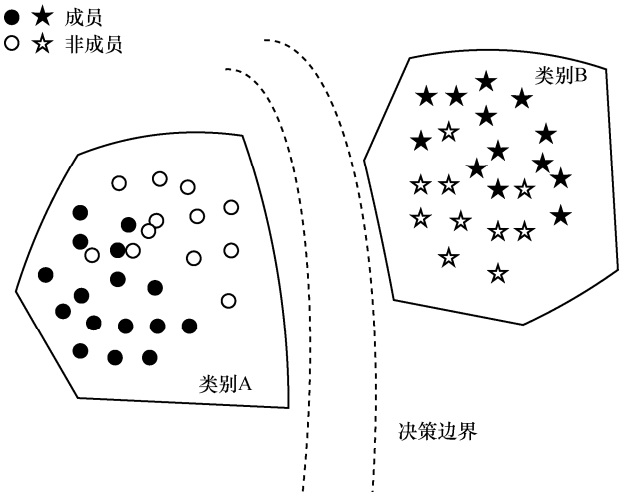


图 3　决策判别

3.2　迁移性分析

本文提出的 PCA-based attack 主要是通过主成分分析技术进行数据降维，在低维流形界面寻找数据的正交映射方向来获取原始数据的对抗样本，再结合快速决策成员推理攻击中基于扰动范畴的算法思想来进行成员推理。高维数据在流模型上的数据映射如图 4 所示，在面对分类图像问题时，将每个类别的数据映射到相应的流形界面，流形上的数据点可以局部地用一个低维向量来表征。对于一个 D 维空间上的样本点 x_0，沿着 d 维空间的流形界面 $H(z)$ 的映射方向进行搜寻，可得到相应的对抗区域和对抗样本。

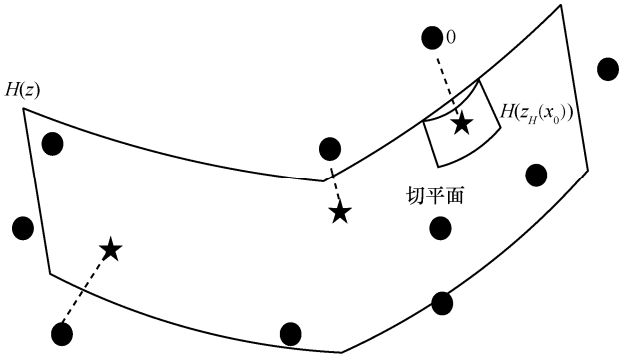


图 4　高维数据在流模型上的数据映射

此外，根据对抗区域的定义，对抗区域中的数据点对所有机器学习模型算法都构成了潜在威胁。由于不同的机器学习模型算法可能具有不同的决策超平面 f_1 和 f_2，因此可以使用这些超平面将对抗区域划分为 2 个子集，即对抗子集 S_{adv} 和常规子集 S_{reg}。如图 5 所示，对抗区域由超平面 f_1 划分得到 S^1_{reg} 和 S^1_{adv}。若该对抗区域又被超平面 f_2 划分，此时将总共得到 4 个区域子集。此时的 S'_{adv} 都被划分为对抗子集，即 $S'_{\text{adv}} = S^1_{\text{adv}} \bigcap S^2_{\text{adv}}$，则表明 2 个决策模型都对 S'_{adv} 中的样本进行错误分类。即 2 个对抗子集的交集中的样本能够在 2 个模型之间传递，该原理说明 PCA-based attack 具有较强的可迁移性。

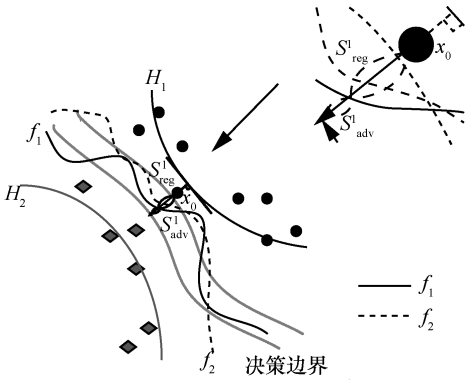


图 5　决策模型的对抗区域

4　仿真实验

为了验证本文提出的 PCA-based attack 的有效性，本文在 3 个真实数据集和一种卷积神经网络模型上进行实验，并与最新攻击进行比较，验证本文攻击的有效性。

4.1　数据与实验参数设置

本文对 CIFAR10[4]、CIFAR100[4]和 GTSRB[4]这 3 个经典的图像数据集进行成员推理实验。首先，基于每个数据集训练 3 组不同数量的数据用于训练模型。另外，由于快速决策成员推理攻击需要多次查询来扰乱数据样本以更改它们的预测标签，因此为基于距离符号梯度的快速决策成员推理攻击设置了查询上限 1×10^5，以进一步研究查询成本对推断性能的影响。为了研究 PCA-based attack 对不同机器学习模型的迁移效果，增添了一组实验数据集 MNIST[3]，且另外部署了 4 组不同架构设置的卷积神经网络 $\{\text{CNN}_7,\text{CNN}_8,\text{CNN}_9,\text{CNN}_{12}\}$ 用于比较算法的迁移。最后，为了进行评估，对 D_{target} 中的数

据进行随机重组，一部分用于训练目标模型 f，即 D_{train}，作为目标模型的成员样本；另一部分 D_{test} 作为非成员样本。评估算法效率时，使用相同大小的集合来最大限度地提高推断的不确定性。

本文实验的源模型是 CNN，模型训练采用 Adam[23]优化器进行优化，其中 epoch=15，batch size=128，learning rate $=1\times10^{-4}$，decay $=1\times10^{-6}$。

由于 AUC 指标考虑了阈值变动的影响，且 ROC 曲线有一个很好的特性：当测试集中的正负样本分布发生变化时，ROC 曲线保持不变。因此，本文实验的评价指标采用 AUC。

4.2 对比攻击方法

为了验证 PCA-based attack 的有效性，本文将其与快速决策成员推理攻击和其他 3 种攻击进行比较，分别为 score-based attack[3-4,7,24-25]、baseline-attack[8]和 boundary-attack[9]。下面对 3 种攻击进行简要介绍。

1) score-based attack。该攻击将攻击转化为一个有监督的二分类问题，利用模拟数据集构建类似目标模型的影子模型，并基于影子模型和目标模型的输出结果训练一个能够判断是否是目标模型训练数据的攻击模型。

2) baseline-attack。该攻击通过数据样本是否被正确分类来进行成员推理。若目标数据被错误分类，则认定该数据为非成员数据，反之为成员数据。具体表达式为

$$\mathcal{A}(x)=\text{sign}(f(x),y)=\begin{cases}-1, f(x)\neq y\\ 1, \text{其他}\end{cases} \tag{17}$$

在实际应用中，不管是模型稳定的算法还是容易过度拟合的算法都容易受到成员推理攻击。

3) boundary-attack。该攻击中，对手不能访问预测置信得分，只能借助目标模型的决策标签来发动攻击。首先利用扰动技术对目标数据点进行决策变动，生成对抗样本；然后计算对抗样本与原始目标之间的变动差异，进而寻找训练数据和测试数据之间的预测差异；最后比较预测差异获得细粒度的成员信号，以实现目标人群的成员推理。

4.3 攻击实验

在攻击的过程中需要解决 2 个主要的问题。1) 在只给定输出标签的黑盒设置中，保证推理精度的同时需要降低访问成本。2) 在访问成本受限的情况下，尽可能消除外在情况带来的影响。

1) 在黑盒设置下的推理性能

首先，为了验证攻击方法在黑盒设置下对目标模型的推理效果，本文在 CNN 模型上对各攻击进行测试，不同攻击在黑盒设置下的推理精度如表 2 所示。

表 2 不同攻击在黑盒设置下的推理精度

数据集	baseline-attack	score-based attack	boundary-attack	fast-attack	PCA-based attack
CIFAR10	0.602	0.707	0.713	0.709	0.687
CIFAR100	0.886	0.921	0.944	0.943	0.923
GTSRB	0.587	0.687	0.724	0.726	0.709

由表 2 可知，大部分的攻击都能实现一定的推理性能。在规模较大的 CIFAR10 数据集（5 000）和 CIFAR100（8 000）数据集中，boundary-attack 和 fast-attack 的推理精度较高，其原因在于两者均基于预测标签获取最优的扰动来区分成员与非成员样本，因此对细粒度的成员信号识别具有较大的影响。而在较小规模的 GTSRB 数据集（600）中，两者攻击精度下降明显，但 fast-attack 依旧维持最优攻击。本文提出的 PCA-based attack 虽然没有得到最优的推理精度，但是在整体上均能保持与 score-based attack 相近的性能。这也验证了 PCA-based attack 能较好地对目标模型进行有力威胁。

本文提出的 fast-attack 不仅在推理精度上取得了不错的效果，在降低成本方面也表现良好，fast-attack 精度随着访问量的变化情况如图 6 所示。

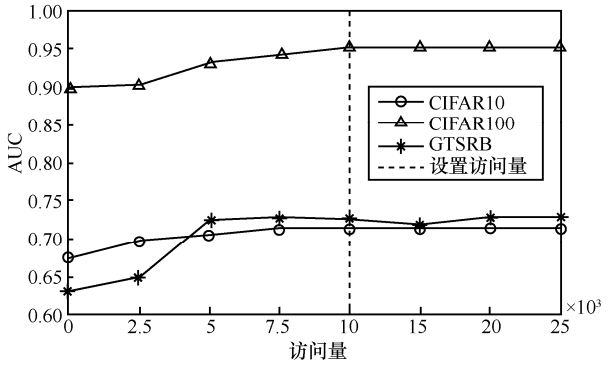


图 6 fast-attack 精度随着访问量的变化情况

在 boundary-attack 中，实验设置访问量为 0～15 000，当访问量的值设置为 10 000 时，在继续增大访问量的情况下，攻击性能不发生明显变化。由图 6 可知，多次随机实验，fast-attack 在限定访问量的情况下，相比 boundary-attack，提前达到最优攻击性能。此外，该算法在 GTSRB 数据集上

收敛速度加倍。因此，本文提出的 fast-attack 在保证推断精度的情况下，降低了模型的访问量成本，甚至在少数数据集上，收敛速度翻倍。

2) 成本受限情况下的迁移能力

从安全的角度来看，可迁移性是攻击的一个重要属性，因为它使敌手能够创建出可以攻击任意目标模型的算法。本文使用文献[26-27]的标准来衡量可迁移性，即由 CNN_7 得出的对抗样本同时又被其他决策模型错误分类所占总体比重来衡量。

为了验证攻击的模型迁移能力，本文在 MNIST 数据集进行了实验，不同算法在 MNIST 数据集的迁移率如表 3 所示。

表 3　不同算法在 MNIST 数据集的迁移率

AUC	攻击	50 000	5 000	500
0.65	boundary-attack	81.63%	51.53%	18.30%
	fast-attack	70.78%	50.02%	12.92%
	PCA-based attack	64.86%	42.78%	21.04%
0.70	boundary-attack	17.74%	3.56%	2.00%
	fast-attack	28.51%	6.14%	1.69%
	PCA-based attack	47.20%	21.12%	8.72%
0.75	boundary-attack	2.55%	—	—
	fast-attack	24.58%	4.53%	1.01%
	PCA-based attack	37.49%	14.79%	4.41%

由表 3 可知，PCA-based attack 的迁移率随推断精度的提升而变大，且明显高于 fast-attack。在推断精度为 0.65 时，PCA-based attack 整体迁移率低于 fast-attack（数据量为 50 000 和 5 000 时），但随着精度的提升，PCA-based attack 远超出其他攻击。实验表明，PCA-based attack 的适应范围更广，攻击效能更强。尽管 PCA-based attack 的推断精度较低，但相比于 fast-attack 需要依赖目标模型来进行推断等决策方法，其不需要利用源模型的任何信息，即可构建性能不错的成员攻击。不同攻击的部署结构如表 4 所示。

由表 4 可知，fast-attack 仅需预测标签即可进行成员推理；而 PCA-based attack 不需要目标模型结构仅需数据分布即可完成推断。相比其他攻击方案，所需条件更少，访问成本更低，更符合实际需求。其中，AUC 结果的设定基于 CIFAR10 模型得到，作为前提条件，便于对比不同攻击方案的部署结构以及攻击性能。

此外，基于主成分的决策边界成员推理攻击中，逻辑判别中的阈值选取尤为重要，实验通过 L2 距离阈值的设定来观察攻击性能。攻击性能随 L2 距离阈值变化如图 7 所示。由图 7 可知，攻击性能随 L2 距离阈值呈凸型变化，在阈值 0.5～1.5 达到峰值。实验表明，要得到一个较优的算法需要选取中间的阈值。

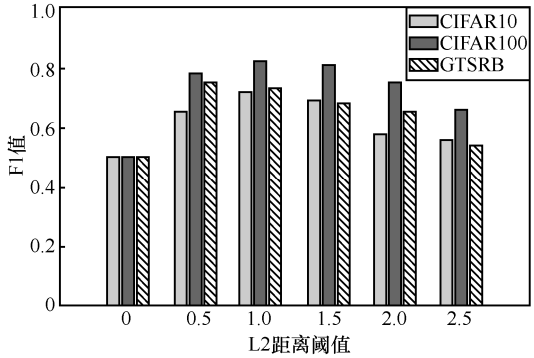


图 7　攻击性能随 L2 距离阈值变化

4.4　有效性分析

4.3 节主要对 PCA-based attack、fast-attack 和其他攻击进行比较，测试了不同场景、不同模型上的推理性能，本节将进一步从抵抗防御角度探究所提方法的有效性。

1) 泛化增强[3-4,28-33]。基于过拟合造成的成员推理攻击，该类方法借助 L1、L2 正则化、随机失活以及数据增强等措施降低模型的过拟合，在提升目标模型预测性能的同时降低数据泄露的风险。

2) 隐私增强[34-36]。差分隐私被广泛用于降低隐

表 4　不同攻击的部署结构（CIFAR10 数据集）

推断类别	攻击	影子模型	机器学习算法	目标模型结构	目标数据分布	置信得分	预测标签	AUC
基于置信度	score-based attack	√或—	√或—	√或—	√或—	√	—	0.707
	baseline-attack	—	—	—	√	—	√	0.602
基于决策	boundary-attack	—	—	—	√或—	—	√	0.713
	fast-attack	—	—	—	—	—	√	0.709
无目标决策	PCA-based attack	—	—	—	√	—	—	0.687

私威胁。该防御技术通过向模型梯度、目标函数添加噪声来防止数据的信息泄露。

3) 置信度扰动[11,37]。以往基于置信度分数的推理攻击能够清晰地呈现成员细粒度信号。因此该类防御旨在改变置信度分数，代表性技术为 MemGuard 和 Adversarial regularization，它们通过改变输出概率分布，使成员与非成员难以区分从而实现防御。

为了验证本文所提攻击的有效性，将不同的成员推理攻击应用于不同的防御技术，实验在 CIFAR10 数据集上使用不同的防御指标参数训练了 3 组目标模型，分别为 Ll($\lambda \in [0.0001, 0.001, 0.005]$)，L2($\lambda \in [0.01, 0.05, 0.1]$) 差分隐私添加的噪声服从高斯分布 $\mathcal{N}(0, \beta), \beta \in [0.1, 0.5, 1.0]$，数据增强通过改变模型训练数据集样本量来验证。

实验表明，在 $\beta = 1.0, \lambda_{L1} = 0.005, \lambda_{L2} = 0.1$ 的情况下，大部分攻击较低了方案的攻击性能，但是损失了目标模型的决策性能，影响模型的实际应用。而 PCA-based attack 并未受影响，是因为该攻击不借助目标模型进行攻击。本文结合实验和理论分析，不同攻击的防御情况如表 5 所示，其中，↓代表攻击性能下降，—代表攻击性能不变。由表 5 可知，在大多数情况下，fast-attack 和 PCA-based attack 都能取得不错的效果，其不仅突破了常见的一些防御技术，甚至目前最优的防御技术 MemGuard 和 Adversarial regularization 都失去了防御效用。因为大部分防御措施主要用于降低模型的过拟合，其针对基于过拟合得到的成员推理攻击能够产生显著效果，但本文攻击借助对抗样本解决了传统成员推理攻击固有的过拟合问题，且目前最优防御技术的原理在于干扰模型的输出置信度。因此，本文提出的成员推理攻击能够规避这些攻击。尽管 fast-attack 和 PCA-based attack 能够规避大多数防御，但是前者难以抵挡差分隐私和 L2 正则化防御，且后者也对数据增强失去效用。这是因为差分隐私通过向目标函数添加噪声干扰了敌手的攻击，而数据增强技术会干扰流模型的形成，进一步影响对抗样本的生成。尽管如此，差分隐私在防御攻击的同时也会干扰模型的效用，难以达到较优的隐私–效用均衡且 L2 正则化在过强的防御干扰下同样会使目标模型失去效用。综上，本文提出的攻击具有较强的稳健性和攻击性。

5 结束语

本文研究了机器学习训练数据集的隐私攻击问题，提出了新的成员推理隐私攻击，即 fast-attack 和 PCA-based attack。前者以低成本快速生成不易感知的对抗样本，从而达到较高精度成员推理。而后者针对 fast-attack 存在的低迁移率问题进行改进，将快速决策成员推理攻击中基于扰动算法与主成分分析技术相结合来进行成员推理，能够在不同模型之间进行高效率迁移。尽管 PCA-based attack 攻击率低于 fast-attack，但相比于 fast-attack 需要依赖目标模型来进行推断等一系列决策算法，其不需要利用源模型的任何信息即可完成成员推理。此外，本文提出的攻击都能对大多数防御的机器学习模型进行攻击，在更严格的对抗模型中实现高精度的成员推理。

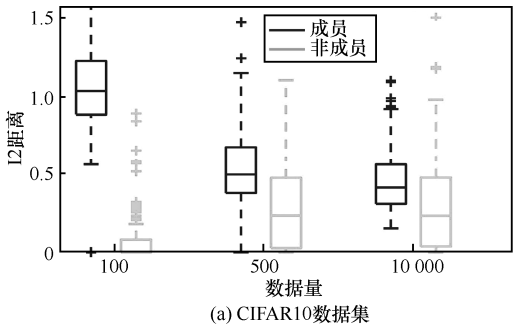
鉴于本文提出的攻击是通过将机器学习的过拟合特性映射到训练集样本与测试集样本的扰动问题中，借助对抗样本来实现成员推理。因此，未来的模型隐私防护工作可在对抗样本的扰动上进行防御工作，进而保护数据的隐私。

附录 1 验证数据的实验分析

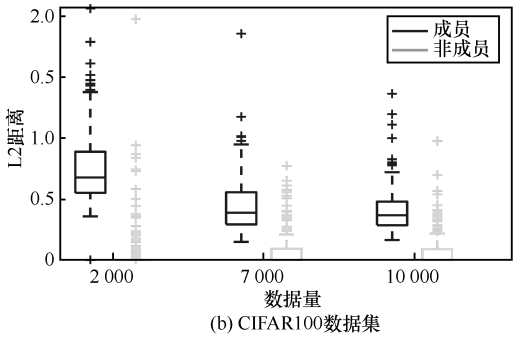
对 CIFAR10、CIFAR100、GTSRB 这 3 组数据的评估数据集进行扰动差异验证，其中，扰动差异通过计算原始数据与扰动数据的 L2 距离得到，结果如图 8 所示。由图 8 可知，成员数据的扰动难度明显大于非成员数据，且随着模型的过拟合程度增大而增大，表明模型的过拟合能够促进成员与非成员样本的细粒度区分，提升攻击性能。

表 5 不同攻击的防御情况

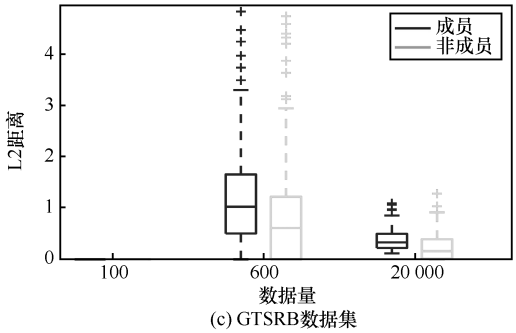
攻击	L1 正则化	L2 正则化	数据增强	差分隐私	MemGuard	Adversarial regularization
baseline-attack	↓	↓	↓	↓	—	—
score-based attack	↓	↓	↓	↓	↓	↓
boundary-attack	↓	↓	—	↓	—	—
fast-attack	↓	↓	—	↓	—	—
PCA-based attack	—	—	↓	↓	—	—



(a) CIFAR10数据集



(b) CIFAR100数据集



(c) GTSRB数据集

图 8　成员与非成员的扰动差异

除了对 MINIST 数据集进行攻击方案的迁移性能验证，还将其扩展到 CIFAR10、ImageNet、GTSRB 数据集，实验结果如表 6、表 7 所示。由表 6、表 7 可知，在小样本数据下，PCA-based attack 的迁移性能表现更优，但在部分大样本数据以及低维数据中，表现欠佳。

表 6　不同攻击在 CIFAR10 数据集的迁移率

AUC	攻击	50 000	5 000	500
0.65	boundary-attack	93.63%	89.53%	52.30%
	fast-attack	90.78%	70.02%	42.92%
	PCA-based attack	77.86%	60.78%	53.04%
0.70	boundary-attack	87.74%	73.56%	37.11%
	fast-attack	86.51%	65.14%	31.69%
	PCA-based attack	62.86%	43.12%	37.72%
0.75	boundary-attack	70.55%	45.4%	19.22%
	fast-attack	71.58%	46.47%	17.01%
	PCA-based attack	42.49%	24.22%	20.95%

表 7　不同攻击在 ImageNet 数据集和 GTSRB 数据集的迁移率

AUC	攻击	ImageNet	GTSRB
0.65	boundary-attack	99.4%	58.53%
	fast-attack	84.1%	53.02%
	PCA-based attack	82.6%	58.78%
0.70	boundary-attack	97.0%	43.56%
	fast-attack	74.4%	36.14%
	PCA-based attack	74.0%	42.12%
0.75	boundary-attack	90.5%	22.7%
	fast-attack	59.9%	21.8%
	PCA-based attack	56.3%	25.89%

参考文献：

[1] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations. [S.l.:s.n.], 2015: 33-47.

[2] MUÑOZ-GONZÁLEZ L, BIGGIO B, DEMONTIS A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2017: 27-38.

[3] SHOKRI R, STRONATI M, SONG C Z, et al. Membership inference attacks against machine learning models[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2017: 3-18.

[4] SALEM A, ZHANG Y, HUMBERT M, et al. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models[C]//Proceedings of 2019 Network and Distributed System Security Symposium. Virginia: Internet Society, 2019: 243-160.

[5] AL-RUBAIE M, CHANG J M. Privacy-preserving machine learning: threats and solutions[J]. IEEE Security & Privacy, 2019, 17(2): 49-58.

[6] MELIS L, SONG C Z, DE CRISTOFARO E, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 691-706.

[7] PYRGELIS A, TRONCOSO C, DE CRISTOFARO E. Knock knock, who's there? membership inference on aggregate location data[C]//Proceedings of 2018 Network and Distributed System Security Symposium. Virginia: Internet Society, 2018: 199-213.

[8] YEOM S, GIACOMELLI I, FREDRIKSON M, et al. Privacy risk in machine learning: analyzing the connection to overfitting[C]//Proceedings of 2018 IEEE 31st Computer Security Foundations Symposium. Piscataway: IEEE Press, 2018: 268-282.

[9] CHOO C A C, TRAMER F, CARLINI N, et al. Label-only membership inference attacks[J]. arXiv Preprint, arXiv: 2007.14321, 2020.

[10] LI Z, ZHANG Y. Membership leakage in label-only exposures[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2021: 880-895.

[11] JIA J Y, SALEM A, BACKES M, et al. MemGuard: defending against black-box membership inference attacks via adversarial examples[C]//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 259-274.

[12] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 739-753.
[13] HAYES J, MELIS L, DANEZIS G, et al. LOGAN: membership inference attacks against generative models[J]. arXiv Preprint, arXiv: 1705.07663, 2017.
[14] LEINO K, FREDRIKSON M. Stolen memories: leveraging model memorization for calibrated white-box membership inference[C]// Proceedings of the 29th USENIX Security Symposium. Berkeley: USENIX Association, 2020: 1605-1622.
[15] LONG Y H, BINDSCHAEDLER V, WANG L, et al. Understanding membership inferences on well-generalized learning models[J]. arXiv Preprint, arXiv: 1802.04889, 2018.
[16] KHALID F, ALI H, ABDULLAH H M, et al. FaDec: a fast decision-based attack for adversarial machine learning[C]//Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-8.
[17] OREKONDY T, SCHIELE B, FRITZ M. Knockoff nets: stealing functionality of black-box models[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 4949-4958.
[18] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.
[19] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[J]. arXiv Preprint, arXiv: 1712.04248, 2017.
[20] CHEN J B, JORDAN M I, WAINWRIGHT M J. HopSkipJumpAttack: a query-efficient decision-based attack[C]//Proceedings of 2020 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2020: 1277-1294.
[21] RIFAI S, DAUPHIN Y N, VINCENT P, et al. The manifold tangent classifier[J]. Advances in Neural Information Processing Systems, 2011, 24(8): 2294-2302.
[22] ZHANG Y G, TIAN X M, LI Y, et al. Principal component adversarial example[J]. IEEE Transactions on Image Processing, 2020, 29: 4804-4815.
[23] KINGMA D, BA J. Adam: a method for stochastic optimization[J]. arXiv Preprint, arXiv: 1412.6980, 2014.
[24] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust You? ": explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1135-1144.
[25] CHEN D F, YU N, ZHANG Y, et al. GAN-leaks: a taxonomy of membership inference attacks against generative models[C]//Proceedings of 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 343-362.
[26] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[J]. arXiv Preprint, arXiv: 1611.01236, 2016.
[27] SIMARD P, VICTORRI B, LE CUN Y, et al. Tangent prop: a formalism for specifying selected invariances in an adaptive network[C]// Proceedings of the 4th International Conference on Neural Information Processing Systems. New York: ACM Press, 1991: 895-903.
[28] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.
[29] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2017: 39-57.
[30] HUI B, YANG Y C, YUAN H L, et al. Practical blind membership inference attack via differential comparisons[J]. arXiv Preprint, arXiv: 2101.01341, 2021.
[31] LI J C, LI N H, RIBEIRO B. Membership inference attacks and defenses in supervised learning via generalization gap[J]. arXiv Preprint, arXiv: 2002.12062, 2020.
[32] SRIVASTAVA N, HINTON G E, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[33] SONG L W, SHOKRI R, MITTAL P. Privacy risks of securing machine learning models against adversarial examples[C]//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2019: 241-257.
[34] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2016: 308-318.
[35] IYENGAR R, NEAR J P, SONG D, et al. Towards practical differentially private convex optimization[C]//Proceedings of 2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2019: 299-316.
[36] RAHIMIAN S, OREKONDY T, FRITZ M. Differential privacy defenses and sampling attacks for membership inference[C]// Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2021: 193.
[37] NASR M, SHOKRI R, HOUMANSADR A. Machine learning with membership privacy using adversarial regularization[C]//Proceedings of 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2018: 634-646.

[作者简介]

彭长根（1963− ），男，贵州锦屏人，博士，贵州大学教授，主要研究方向为隐私保护、密码学和大数据安全等。

高婷（1995− ），女，江西吉安人，贵州大学硕士生，主要研究方向为隐私保护、成员推理等。

刘惠篮（1988− ），女，贵州贵阳人，博士，贵州大学副教授，主要研究方向为复杂数据分析、稳健回归、高维数据建模和统计计算。

丁红发（1988− ），男，河南南阳人，博士，贵州大学在站博士后，贵州财经大学副教授，主要研究方向为隐私保护和大数据安全。